Note 4 - Singular Value Decomposition

Agnibho Roy

Fall 2020

1 The SVD Theorem

Motivation. SVD is a general factorization that extends beyond only symmetric matrices. Very powerful properties of matrices can be derived from this such as the spectral norm, subspace condition numbers, PCA, and least squares for overdetermined and underdetermined systems.

Key Idea. We already have a nice factorization for $A_s \in \mathbb{S}_n$ as we can recall from the spectral theorem, but what about general matrices? Instead, we consider A^TA for $A \in \mathbb{R}^{m\times n}$, which we can verify to be symmetric and PSD. We can further verify that for $v \in \mathbb{R}^n$,

$$\|Av\|_2 = \|\sqrt{A^TA}v\|_2$$

So $\sqrt{A^TA}$ is nothing but an orthogonal transformation. To transform back, we can take some rotation matrix P where $P\sqrt{A^TA} = A$. SVD just factorizes this later equation into:

$$P\sqrt{A^TA} = (PV)\sqrt{\Lambda}\left(V^T\right) = U\Sigma V^T$$

Which is the infamous equation that we all know. Graphically, we can see SVD as an orthogonal transformation to a new basis (V^T), then a scaling by the singular values (Σ), and then another orthogonal transformation (U). We now formalize these matrices.

Lemma 1. A^TA and AA^T have the same nonzero eigenvalues and hence the same rank r. They are also symmetric and PSD.

Theorem 2 (SVD Compact Form)**. Any matrix $A \in \mathbb{R}^{m\times n}$ can be factored as $A = U_r\Sigma V_r^T$ for orthogonal $U_r \in \mathbb{R}^{r\times m}$, $V_r^T \in \mathbb{R}^{n\times r}$, and diagonal $\Sigma \in \mathbb{R}^{m\times n}$.**

Proof. Let us consider A^TA's eigenvectors corresponding to *nonzero* eigenvalues (i.e. for v_i where $i = 1\dots r$) and hit the equations by A on the left:

$$A^TAv_i = \lambda_i v_i$$
$$(AA^T)Av_i = \lambda_i Av_i$$

So this means that Av_i are the eigenvectors of AA^T, and we can show that they are actually mutually orthogonal. Remember by the spectral theorem $A^TA = V\Lambda V^T$ for orthonormal V.

$$(Av_i)^T(Av_j) = v_i^T(A^TAv_j) = \lambda_i v_i^Tv_j = \lambda_i \text{ if } i = j \text{ and } 0 \text{ otherwise}$$

So to make the matrix of $U = AV$ orthonormal, we can normalize by defining: $u_i = \frac{Av_i}{\sqrt{\lambda_i}}$. So for each i, j, we now have:

$$u_i^TAv_j = \frac{1}{\sigma_i}v_i^TA^TAv_j = \frac{\lambda_i}{\sigma_i}v_i^Tv_j = \sigma_i \text{ if } i = j \text{ and } 0 \text{ otherwise}$$

Putting it all together by stacking up all u_i and v_i

$$U_rAV_r^T = \Sigma \implies A = U_r\Sigma V_r^T$$

□

Lemma 3. $A^TAv_i = 0 \implies Av_i = 0$ for $i = r+1 \dots n$.

Proof. Suppose for contradition that $A^TAv_i = 0$ and $Av_i \neq 0$. Then $A^T(Av_i) = 0$ means that $Av_i \in \mathcal{N}(A^T) = \mathcal{R}(A)^T$ (Fundamental Theorem of Linear Algebra), which is impossible because $Av_i \neq 0$ means that $Av_i \in \mathcal{R}(A)$.

Theorem 4 (SVD Full Form)**. We can extend matrices $U \in \mathbb{R}^{r\times m} \to U \in \mathbb{R}^{n\times m}$, $V^T \in \mathbb{R}^{n\times r} \to U \in \mathbb{R}^{n\times m}$, and Σ to be block diagonal with 0 matrices with appropriate dimensions.**

Proof. The full form of V can easily be found by taking all vectors v_i for $i = 1 \dots n$ instead of just $i = 1 \dots r$. By lemma 3, we know that $Av_j = 0$ for $j = r+1 \dots n$. So if we just extend $u_i \dots u_r$ to create an orthogonal basis for $\mathbb{R}^m$ to $u_1 \dots u_m$ using Gram Schmidt Orthogonalization, then we know that:

$$u_i^T Av_j = 0 \text{ for } i = 1 \dots m \text{ and } j = r+1 \dots m$$

This leads to the augmented $\tilde{\Sigma}$ with the following dimensions because for the specific combination of vectors above, the all map to 0:

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{bmatrix}$$

This leads to the full form $U\tilde{\Sigma}V^T$. □

2 SVD Properties

There are some key properties to note about the SVD, mainly what each of the orthogonal matrices actually explain about the matrix A. They can be summarized into the following five points:

1. The rank of A can be determined by the number of *nonzero* eigenvalues in $\tilde{\Sigma}$.
2. $\{u_1 \dots u_r\}$ form an orthonormal basis for $\mathcal{R}(A)$
3. $\{u_{r+1} \dots u_m\}$ form an orthonormal basis for $\mathcal{N}(A^T)$
4. $\{v_1 \dots v_r\}$ form an orthonormal basis for $\mathcal{R}(A^T)$
5. $\{v_{r+1} \dots v_n\}$ form an orthonormal basis for $\mathcal{N}(A)$

Now the geometric interpretation should be much clearer from the beginning of this section in terms of *exactly* which orthonormal bases we are moving into and from prior to and after scaling by Σ. We now look at the euclidean balls that are formed by the linear transformation by A.

Theorem 5 (Image of Ax using SVD)**. The *image* of the unit ball $\mathcal{B}^n = \{x \in \mathbb{R}^n : x^Tx \leq 1\}$ after translating by A is the set $A \cdot \mathcal{B}^n = \{y \in \mathcal{R}(A) : y^TU_r\Sigma^{-2}U_r^Ty \leq 1\}$.**

Theorem 6 (Preimage of Ax using SVD)**. The *preimage* of the unit ball $\mathcal{B}^n = \{x \in \mathbb{R}^n : x^Tx \leq 1\}$ under A, i.e. all the vectors that map to the unit ball after being translated by A, is the union of the affine translates of $N(A)$ by the vectors in $\mathcal{R}(A^T)$ lying in the ellipsoid $\{V_rz : z^R\Sigma^2z \leq 1\}$**

Theorem 7 (Singular Values of Matrix Operations)**. We can write $A^T = V\tilde{\Sigma}^TU^T$ and so A^T and A have the same singular values. Moreover, for $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times m}$, AB and BA have the same nonzero eigenvalues with same algebraic and geometric multiplicities.**

Proof. The first part follows directly from lemma 1. □

3 Matrix Norms and the Psuedoinverse

We already have an idea for how to take norms of matrices. We first introduced the Frobenius norm, which you can recall is given by:

$$\|A\|_F^2 = \sum_i \sum_j a_{ij}^2 = \text{Trace}(A^T A) = \sum_i \lambda_i(A^T A) = \sum_i \sigma_i^2$$

Other examples of matrix norms are l_p induced norms, which gives a measure of the largest amount by which a matrix "scales up" a vector. A commonly used one is the l_2 induced norm:

Definition 8 (L_2 induced norm)**.** Is given by:

$$\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\lambda_{\max}(A^T A)}$$

Another form of a matrix norm is the nuclear norm, which you can show as an exercise is a proper norm.

Definition 9 (Nuclear Norm)**.** For $r = rank(A) \geq 1$, the nuclear norm is given by:

$$\|A\|_* = \sum_{i=1}^{r} \sigma_i$$

Whenever we cannot find the inverse of a matrix, we can use something called a psuedoinverse that almost "acts" as an inverse. We define it as follows:

Definition 10 (Moore-Penrose Pseudoinverse)**.** If we write $A = U_r \Sigma V_r^T$ from the compact SVD theorem, then we define $A^\dagger \in \mathbb{R}^{m \times n}$ as:

$$A^\dagger = V_r \Sigma^{-1} U_r^T$$

You can think of $A^\dagger$ as some transpose space that is "more aligned to exhibiting properties of some inverse", as you will see from the following properties:

1. $\mathcal{R}(A^\dagger) = \mathcal{R}(A^T)$ and $\mathcal{N}(A^\dagger) = \mathcal{N}(A^T)$.
2. $AA^\dagger = U_r U_r^T$ and $A^\dagger A = V_r V_r^T$.
3. If $r = n$, then we have that $A^\dagger A = I_n$ because this means that $A^T A$ is of full rank and can be inverted. In fact in this case, $A^\dagger = (A^T A)^{-1} A^T$
4. Similarly, if $r = m$, $AA^\dagger = I_m$ and AA^T will be of full rank and can be inverted. In this case, we have that $A^\dagger = A^T (AA^T)^{-1}$.
5. In general, $AA^\dagger A = A$ and $A^\dagger A A^\dagger = A^\dagger$.

In the next note we discover how to use this to solve least-square problems.